SUPPLEMENTARY INFORMATION

Survey of 800+ datasets from human tissue and body fluid reveals XenomiRs are likely artifacts

Wenjing Kang[1†], Claus Heiner Bang-Bertelsen[2,3,4†], Anja Holm[5], Anna Houben[6,7], Anne Holt Müller[8], Thomas Thymann[9], Flemming Pociot[2,10,11], Xavier Estivill[6,7], Marc R. Friedländer[1*]

[1]Science for Life Laboratory, Department of Molecular Biosciences, The Wenner-Gren Institute, Stockholm University, S-10691 Stockholm, Sweden. [2]Center for Non-Coding RNA in Technology and Health, University of Copenhagen, Copenhagen, Denmark. [3]Department of Diabetes Biology, Novo Nordisk, Måløv, Denmark. [4]National Food Institute, Technical University of Denmark, Søborg, Denmark. [5]Molecular Sleep Laboratory, Department of Clinical Biochemistry, Rigshospitalet, Glostrup, Denmark. [6]Bioinformatics and Genomics Programe, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, 08003 Barcelona, Spain. [7]Universitat Pompeu Fabra (UPF), 08002 Barcelona, Spain. [8]Department of Clinical Experimental Research, Glostrup Research Institute, Rigshospitalet, Glostrup, Denmark. [9]Department of Human Nutrition, University of Copenhagen, Frederiksberg, Denmark. [10]Department of Paediatrics, Herlev Hospital, University of Copenhagen, Copenhagen, Denmark. [11]Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.

† The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

* To whom correspondence should be addressed. Tel: +46 737121558 Email: marc.friedlander@scilifelab.se

Contents:

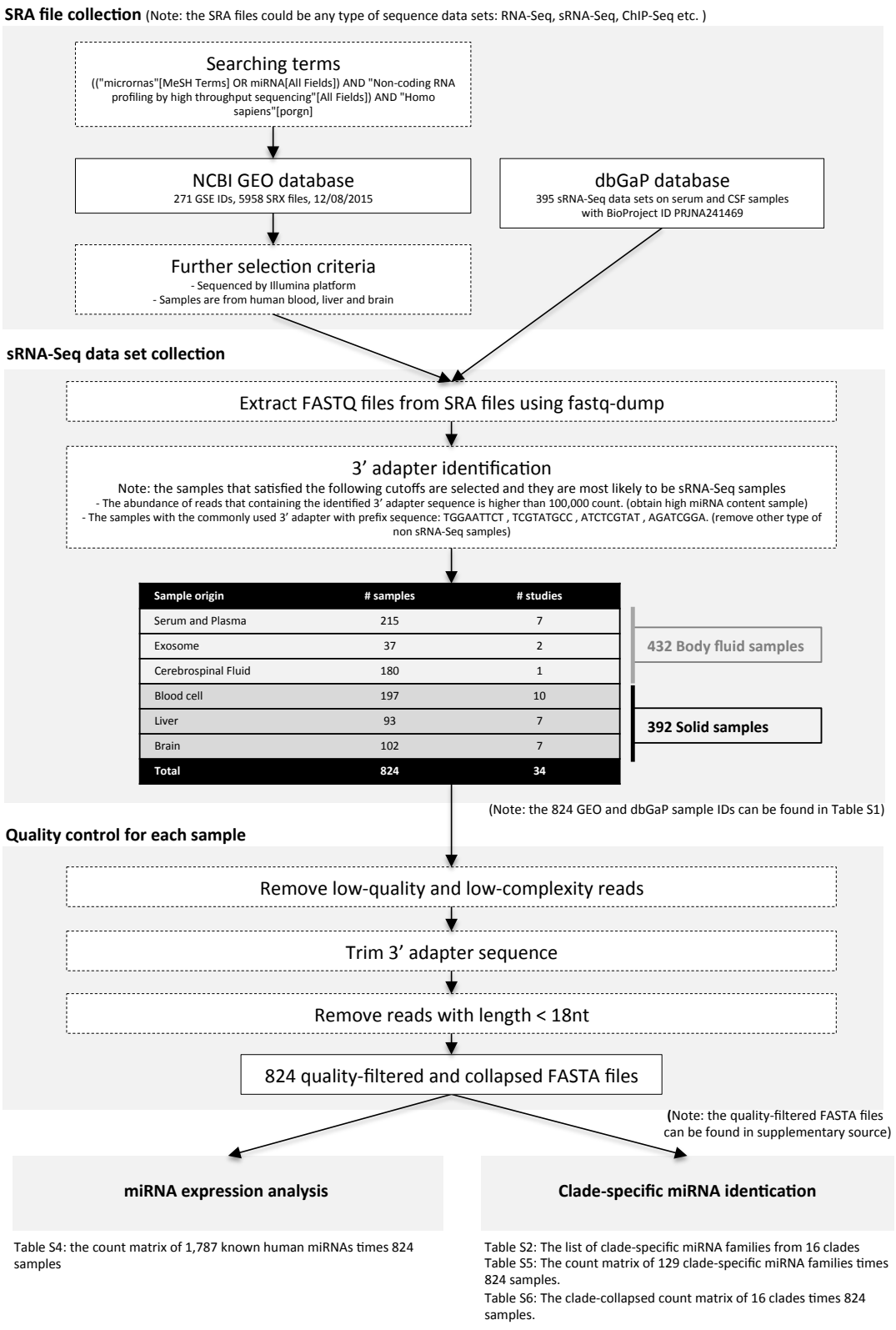

Figure S1, data collection and processing workflow.

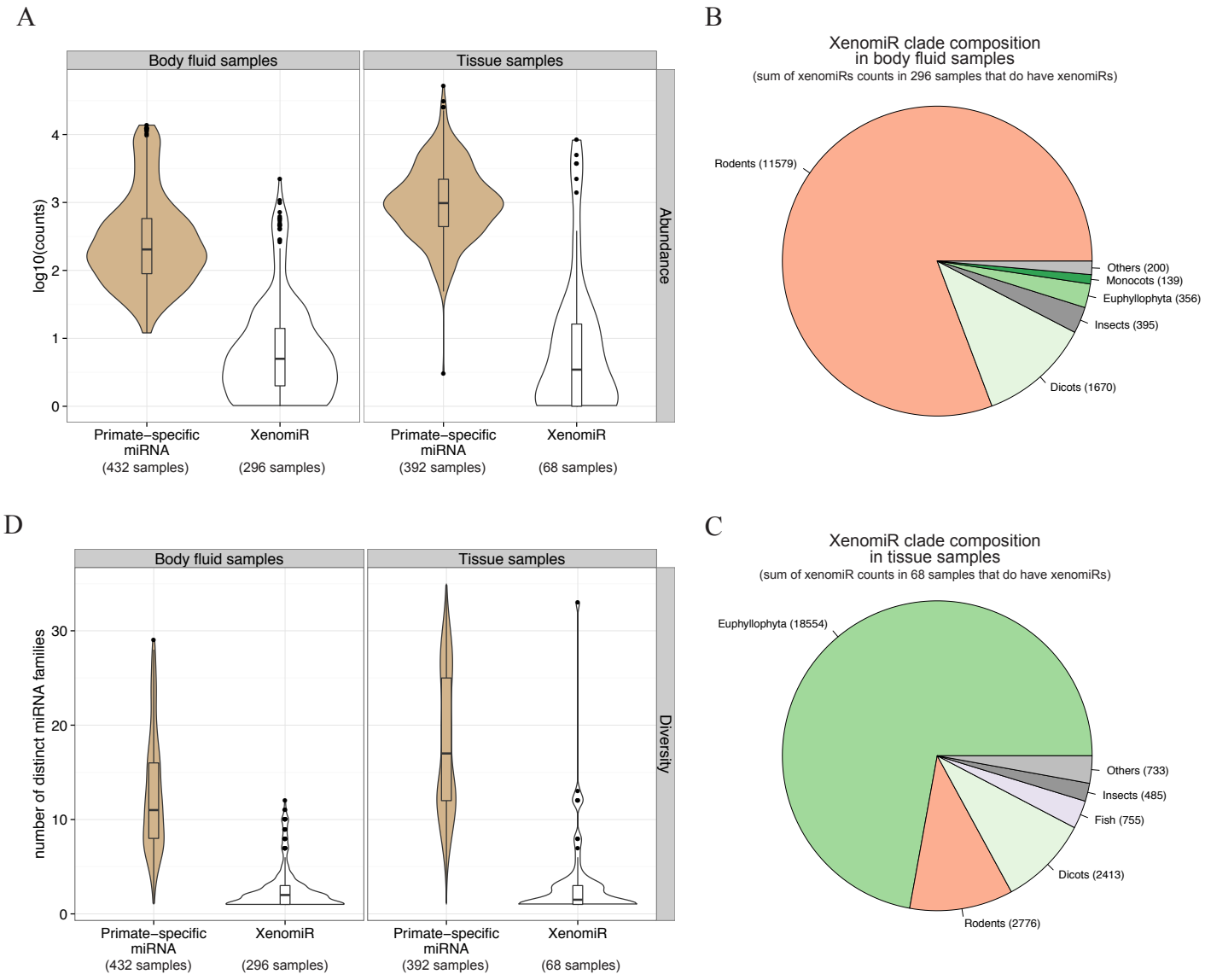

Figure S2, abundance (A-C) and diversity (D) of primate-specific miRNAs and xenomiRs in body fluid samples and tissue samples. The primate-specific miRNAs are more abundant in tissue samples than in body fluid samples ($p < 0.001$, two tail Wilcoxon rank sum test). The xenomiR abundances are comparable in body fluid samples and tissue samples ($p = 0.278$, two tail Wilcoxon rank sum test) (A). The most abundant xenomiRs in body fluid samples are orignated from rodents. The most abundant xenomiRs in tissue samples are orignated from euphyllophyta (B-C). The primate-specific miRNAs are more diverse in tissue samples than in body fluid samples ($p < 0.001$, two tail Wilcoxon rank sum test). The xenomiRs are same diverse (in terms of the number of distinct miRNA families) in body fluid samples and tissue samples ($p = 0.870$, two tail Wilcoxon rank sum test) (D).

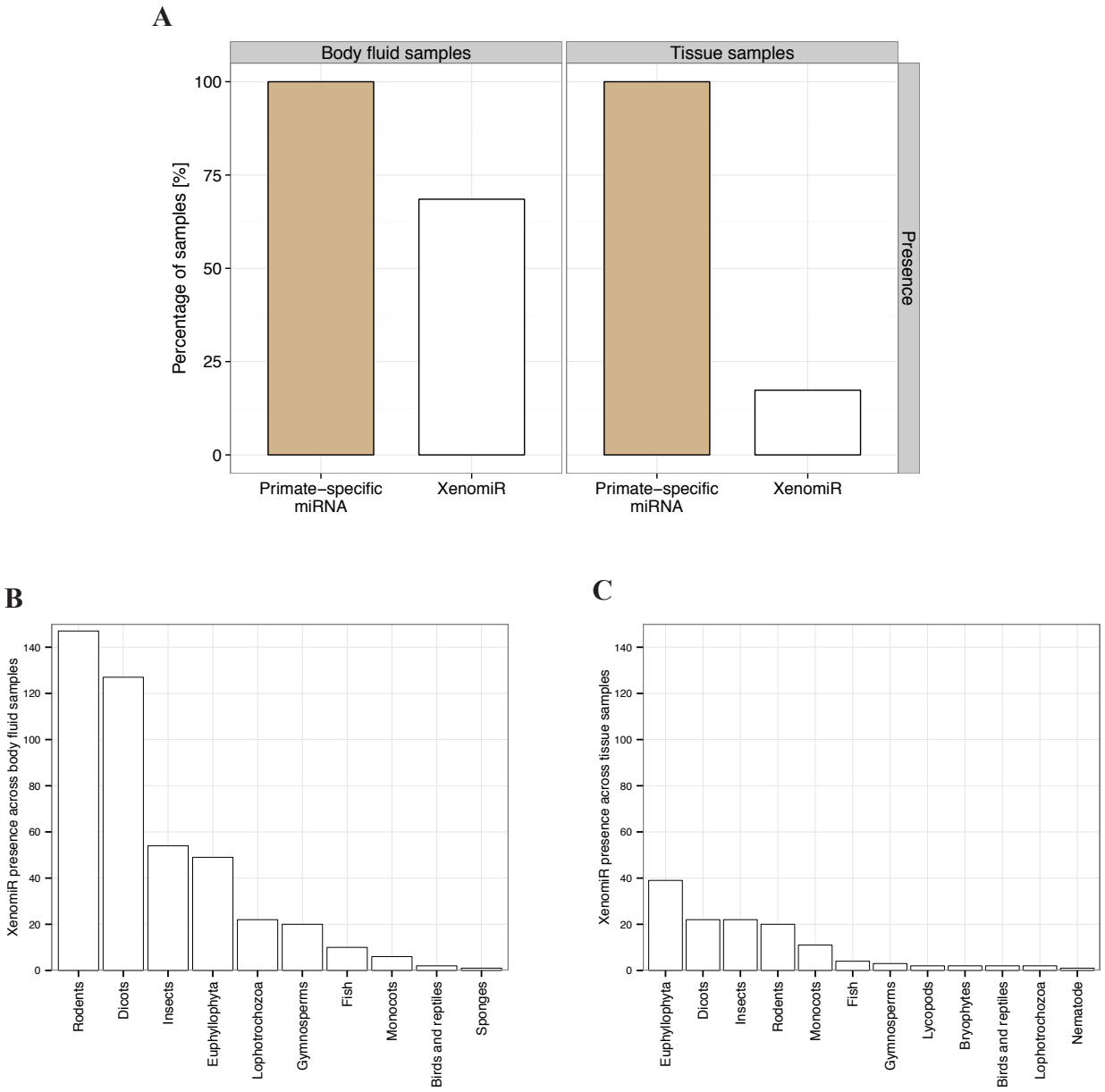

Figure S3, presence of xenomiRs in 824 human samples. The xenomiRs are in general more present in body fluid samples compared to tissue samples. As expected, the primate-specific miRNAs are detected in all human samples (n = 824). The xenomiRs are detected in 296 (68.5%) of 432 body fluid samples and in 68 (17.3%) of 392 tissue samples (A). The rodent, dicot, insect and euphyllophyta xenomiRs are dominate in body fluid samples. For instance, rodent xenomiRs are present in 147 of body fluid samples (B). The euphyllophyta, dicot, insect and rodent xenomiRs are dominate in tissue samples (C).

Figure S4, statistic tests on presences, abundances and diversities of primate miRNAs and xenomiRs

id	The statement	Statistic test	H_0 hypothesis	p-value★
P1.1	XenomiRs are **present** at a lower fraction of blood cells (7%) compared to liver (26%) and brain (30%). XenomiR presence in liver and brain are comparable.	Partition likelihood-ratio Chi-squared test	XenomiR presences are same in blood cells, liver and brain (rejected); XenomiR presences are same in liver and brain (not rejected); XenomiR presences are same in liver/brain versus blood cell (rejected).	< 0.001, 0.477, < 0.001
P1.2	XenomiRs are **present** at a higher fraction of health brain (40%) compared to liver (26%) and blood cell (7%).	Partition likelihood-ratio Chi-squared test	XenomiR presences are same in blood cells, liver and health brain (rejected); XenomiR presences are same in liver and health brain (rejected); XenomiR presences are same in liver versus blood cell (rejected).	< 0.001, 0.09, < 0.001
p2	XenomiRs are **overrepresented** in older studies (present in 88% samples) from before the year 2013 relative to the newer studies (12%).	Chi-squared test	XenomiR presences are same in older and recent studies (rejected).	< 0.001
p3	The older studies are more likely to contain xenomiRs from multiple clades than were the recent ones.	Wilcoxon rank sum test	XenomiRs diversities are same in older and recent studies (rejected).	< 0.001
p4	Primate-specific miRNAs are less **abundant** in body fluid samples than in tissue samples (count median and interpercentile range: 204 (90-580) vs. 979 (441-2195))	Wilcoxon rank sum test	The median primate-specific miRNA counts are same in body fluid and tissue samples (rejected)	< 0.001
p5	Primate-specific miRNAs are less **diverse** in body fluid samples than in tissue samples (diversity median: 11 (8-16) vs. 17 (12-25))	Wilcoxon rank sum test	The median of distinct primate-specific miRNA families are same in body fluid and tissue samples (rejected)	< 0.001
p6	XenomiRs are **overrepresented** in body fluid samples (69%) compared to tissue samples (17%).	Chi-squared test,	XenomiR presences are same in body fluid and tissue samples (rejected)	< 0.001
p7	The **presences** of euphyllophyta, insect, lophotrochozoa and rodent xenomiRs are driven by study.	Permutation test (n = 1000)	XenomiRs from each clade are randomly present across studies. The hypothesis was tested on 6 clades that have the xenomiRs present in more than 10 samples: Dicots, Euphyllophyta, Gymnosperms, Insects, Lophotrochozoa, Rodents	0.821, 0.040, 0.499, < 0.001, < 0.001, < 0.001
p8	XenomiRs are comparable **abundance** in body fluid and tissue samples that do contain xenomiRs (abundance median: 5 (2-14) vs. 3.5 (1-16))	Wilcoxon rank sum test	The median xenomiR counts are same in body fluid and tissue samples that did contain xenomiRs (not rejected)	0.278
p9	XenomiRs are comparable **diversity** (in terms of distinct miRNAs) in body fluid and tissue samples that do contain xenomiRs (diversity median:2 (1-3) vs. 1.5 (1-3))	Wilcoxon rank sum test	The median of distinct xenomiR families are same in body fluid and tissue samples that do contain xenomiRs (not rejected)	0.870
p10	XenomiR **presence**, **diversity** and **abundance** are comparable in serum and CSF samples that from the study PRJNA241469 and do contain xenomiRs. (Presence: 64% vs. 70%, Diversity: 2 (1 to 3) vs. 2 (1 to 3), Abundance: 4 (2 to 10.5) vs. 5(2 to 11.8))	Chi-squared test, Wilcoxon rank sum test, Wilcoxon rank sum test	XenomiR presences are same in serum and CSF samples (not rejected); The median of distinct xenomiR families are same in serum and CSF samples (not rejected); The median xenomiR counts are same in serum and CSF samples from the study PRJNA241469 (not rejected).	0.312, 0.080, 0.569
p11	XenomiRs originated from rodent and dicot clades (each in around 32% of samples) are more **present** than that from insect and euphyllophyta clades (12%) and the rest clades (1%) in body fluid samples.	Partition likelihood-ratio Chi-squared test	XenomiR presences are same in 15 exogenous clades (rejected). Rodents vs. dicots; insects vs. euphyllophyta; lophotrochozoa vs. gymnosperms; fish vs. monocots …	< 0.001, 0.144, 0.600, 0.751, 0.311 …
p12	Rodent xenomiRs are more **abundant** than dicot and than insect and euphyllophyta xenomiRs in body fluid samples that do contain xenomiRs. (Abundance: 5 (2 to 21) vs. 3 (1 to 6) vs. 2 (1 to 2) and 1(1 to 2))	Wilcoxon rank sum test	The median counts of insect and euphyllophyta xenomiR counts are same (not rejected); The median counts of insect/euphyllophyta and dicot xenomiR are same (rejected); The median counts of dicot and rodent xenomiR are same (rejected).	0.184, 0.004, < 0.001
p13	The samples are grouped by studies. The significantly clustered studies are PRJNA241469, GSE59565, GSE58410, GSE45722,.	Permutation test on hierarchical clustering (n= 1000)	The hierarchical clustered samples are randomly distributed across studies. This hypothesis was tested by each study: PRJNA241469, GSE71579, GSE49035, GSE52981, GSE53439, GSE59565, GSE58410, GSE45722.	< 0.001, 0.154, 0.194, 0.753, 0.977, 0.021, 0.013, < 0.001

There are three aspects of comparisons: i) **presence**, the miRNA occurrences across samples; ii) **abundance**, we use the median and interpercentile range to describe abundance; iii) **diversity,** indicated by the number of distinct miRNA families. ★ the significance threshold is set at 0.01, except for the permutation test with the threshold at 0.05.

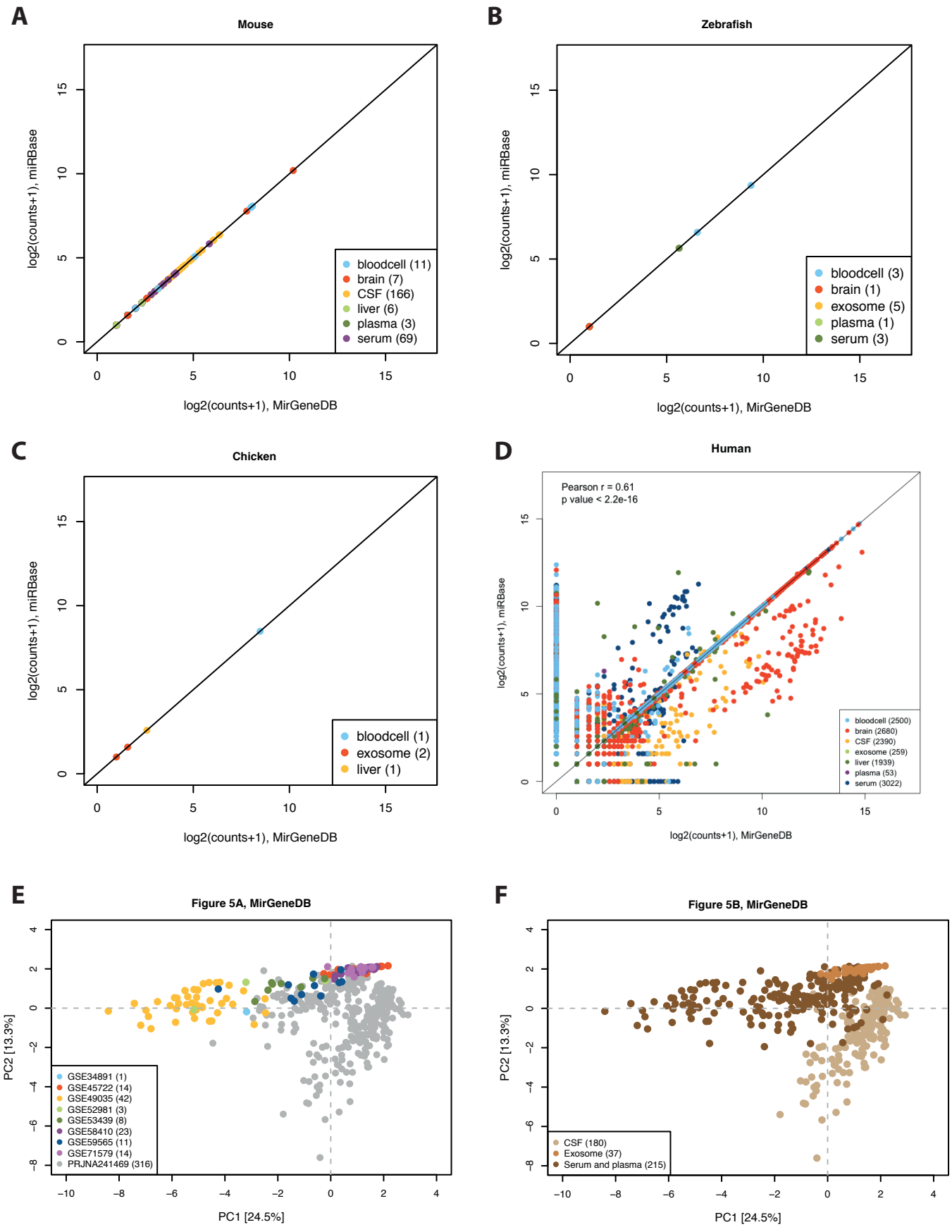

Figure S5, correlation of rodent (A), fish (B), birds (C) and primate (D) specific miRNAs reported using miRBase and the highly curated MirGeneDB. Mouse, zebrafish, chicken and human miRNA sequences from the two databases are used respectively as reference to generate these plots. Each data point represents the read count of one clade-specific miRNA family in one data set. The subplot E and F are regenerated plots of Figure 5A and 5B. Instead of using miRBase as reference, MirGeneDB is used as reference.

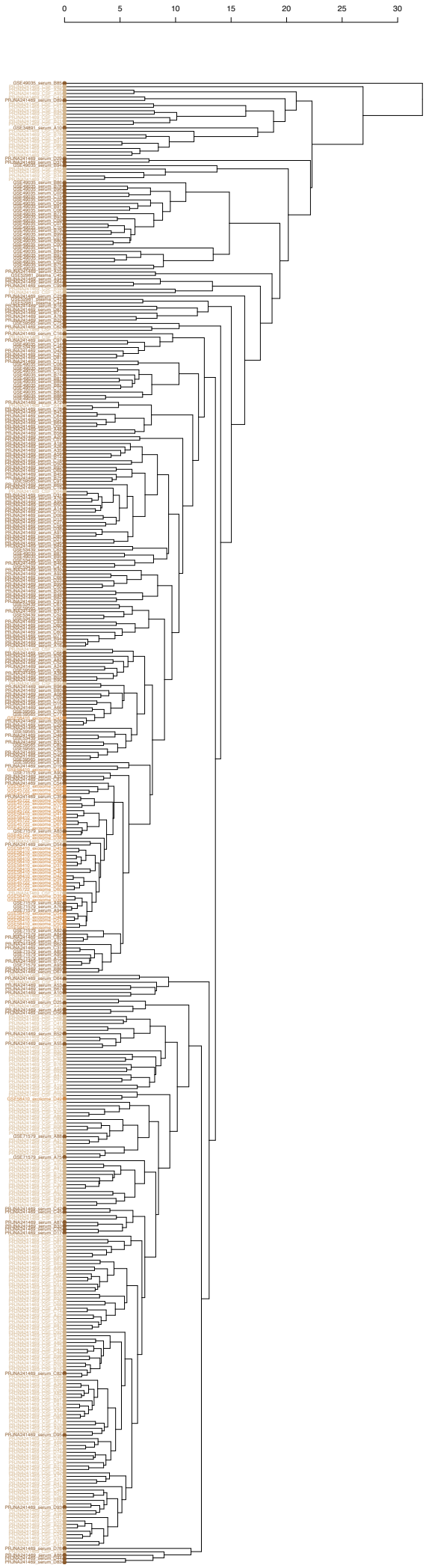

Figure S6, the hierarchical clustering of 432 body fluid samples based on the RPM normalized and $Log_{10}(n_{r,s}+1)$ transformed count matrix of primate-specific miRNA families in Table S5 using R with euclidean distance and complete-linkage. The samples with similar miRNA composition tend to group together. The samples are colored based on tissue type as same as Figure 5B. The serum and plasma, exosome and CSF samples are in dark brown, brown and light brown respectively. The samples are in general grouped by tissue type.

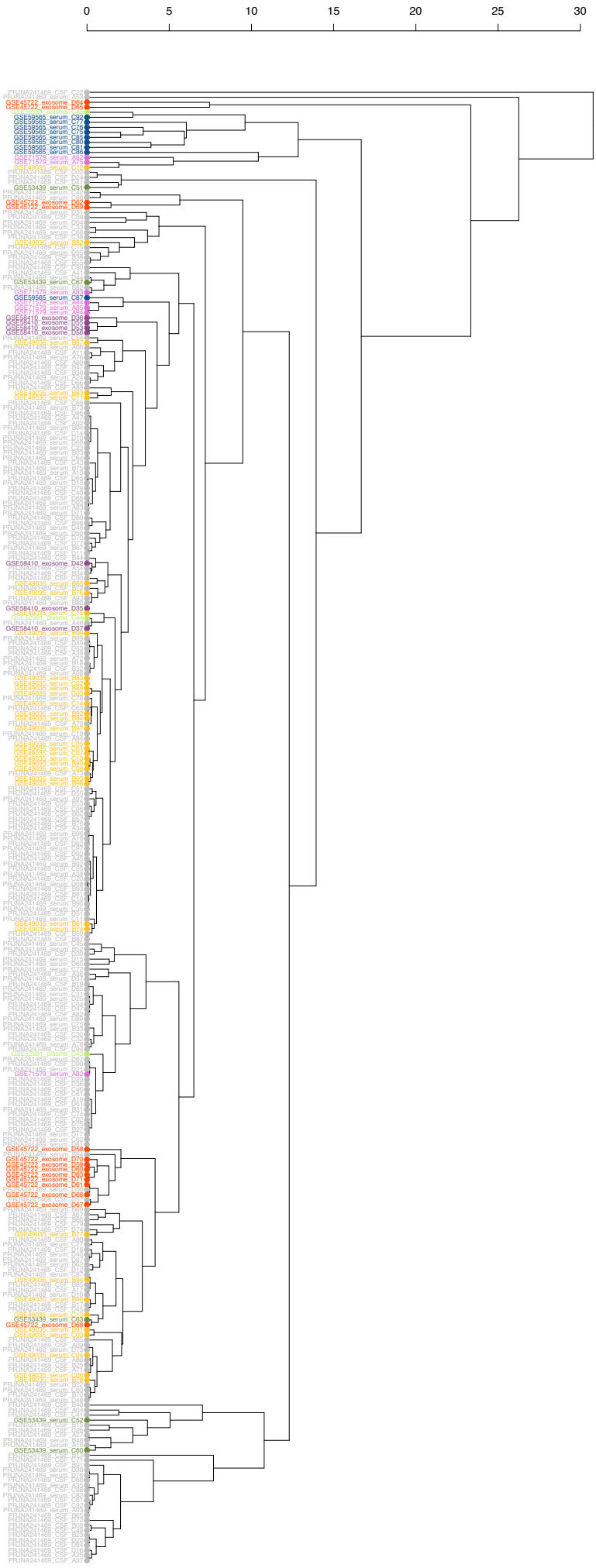

Figure S7, the hierarchical clustering of 432 body fluid samples based on the RPM normalized and $Log_{10}(n_{r,s}+1)$ transformed count matrix of xenomiR families in Table S6 using R with euclidean distance and complete-linkage. 296 out of 432 samples do have xenomiRs. After removing 4 outliers, GSE34891_serum_A10, GSE59565_serum_C83, PRJNA241469_CSF_C22 and PRJNA241469_CSF_B13, 292 samples are showed in the plot. The samples are colored based on studies (GSE accession number) as same as Figure 4C. It looks like the samples from same study tend to group together in exception of samples with BioProject ID PRJNA241469.

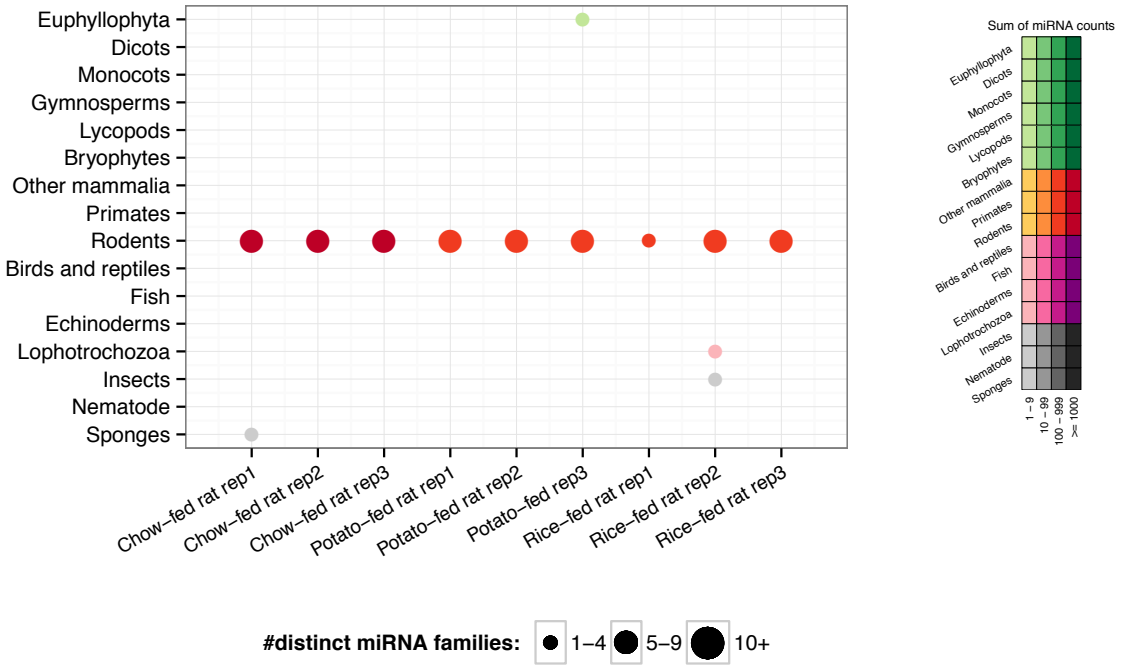

Figure S8, presence and abundance of xenomiRs in chow, potato and rice fed rat samples. As expected, rodent-specific miRNAs are present and abundant in all rat samples. The plant specific miRNAs that originate from euphyllophyta are detected in one of the potato-fed rat sample.

Supplementary source

The 508 FASTA files with the reads that passed the quality control can be downloaded at https://figshare.com/s/36e1c536a573011e4248. The 316 samples from PRJNA241469 are not included due to the controlled-access portion of dbGaP. For each FASTA file, we collapsed identical reads into a single sequence with recording the read count in the ID field. For example, a read from sample "GSE34891_serum_A10.fa" has id ">A10_0_x262795" and sequence "TGAGGTAGTAGTTTGTGCT". The number 262795 after the "x" represents the read count of the sequence. The reads in each FASTA file are sorted based on abundance.